DOCUMENT RESUME

ED 218 342 TM 820 383

AUTHOR Sigmon, Gary; And Others
TITLE Use of Judgmental Procedures by Groups of Raters to Set Minimum Competency Standards.
PUB DATE Apr 81
NOTE 17p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Los Angeles, CA, April 14-16, 1981).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Standards; College Entrance Examinations; Comparative Analysis; *Cutting Scores; Higher Education; *Minimum Competency Testing; Rating Scales; Verbal Tests
IDENTIFIERS Angoff Methods; Ebel Method; Missouri College English Test; Nedelsky Method; *Standard Setting

ABSTRACT

Three questions were addressed by this study: (1) Do the Ebel, Nedelsky, and Angoff methods of setting minimum competency standards produce different cutting scores for the same test? (2) When utilized by raters of different competencies, do these methods result in different Standards? and (3) Are the standards set by the different methods the same across groups of raters? Each rater in each of three groups of raters was asked to rate each of the items in the Missouri College English Test using the three methods. After all rating forms were returned, cutting scores for each of the 15 raters and each of the three groups of raters were determined for each of the three methods. It was concluded that: (1) when utilized to set standards on the same test, the Ebel, Nedelsky, and Angoff methods produce different standards; (2) when these methods are combined, groups of raters set standards that do not differ; and (3) high school teachers using the Nedelsky procedure set the lowest standards. (Author/BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED218342

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

USE OF JUDGMENTAL PROCEDURES BY GROUPS OF RATERS TO SET MINIMUM COMPETENCY STANDARDS

Gary Sigmon, Gerald Halpin,
Glennelle Halpin, and Ellen Sides
Auburn University

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

G. L. Sigmon

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 820 383

Paper presented at the meeting of the National Council on Measurement in Education, Los Angeles, April 1981.

Use of Judgmental Procedures by Groups of Raters
To Set Minimum Competency Standards

Simply stated, a standard, as defined by Andrew and Hecht (1976), refers to the pass-fail level on a particular test. Popham (1978) described standard as "a measure of what [student performance] is adequate for some purpose" (p. 2). Other terms used to describe standard include cut-off score, cutting score, mastery score, passing score, level of minimal competence, and criterion score (Livingston & Zieky, 1978; Linn, 1978; Glass, 1978). A particularly relevant description of the term cut-off score was proposed by Hambleton (1978): "a point on a test score scale that is used to sort examinees into two categories that reflect different levels of proficiency relative to a particular objective measured by a test" (p. 279). Defining the word is no problem but choosing the best method to use when establishing standards for competency assessment in education provides educators a disturbing dilemma. Pressures from both internal and external sources force educators to set standards for performance that will identify minimally competent students in nearly all subject areas. However, general consensus on how these standards are to be set does not exist.

Various authors have categorized and described methods of standard setting (Hambleton, Powell, & Eignor, 1979; Livingston & Zieky, 1978; Popham, 1978; Meskauskas, 1976; Buck, 1977). A commonly cited category of methods in the literature includes the judgmental approaches, such as those described by Ebel (1972), Nedelsky (1954), and Angoff (1971). Judgmental methods, as opposed to normative and criterion methods, do not require the use of examinee data to set the standard. With judgmental methods of standard setting, raters have access to the individual items on the test for which standards are to be set, and decisions are made relative to the

items rather than to external criteria. Even when focusing on the judgmental approach to standard setting, conscientious educators, who realize the significance of their decisions, still must select the method they will use. They could have more confidence in their choice if similar standards resulted from the various methods, but is such a realistic expectation? Andrew and Hecht (1976) asked if different methods employed by the same and different groups produced different standards. They found that use of the Nedelsky and Ebel methods did indeed produce different standards, but there was relative stability within methods across groups. Skakun and Kling (1980) also found that utilization of the Nedelsky method and a modified version of the Ebel approach produced different standards. Koffler (1980) found inconsistencies in the cut-off scores obtained by using the Nedelsky procedure and the contrasting groups method.

These studies, with their select samples, measures, and standard-setting procedures, are just the beginning of needed research in an area with such widespread applicability and significant consequences. More evidence regarding the comparability, or lack of it, among the various standard-setting methods utilizing different measures and groups is needed.

Background

Students seeking admission to teacher education at the large state university where this study was conducted must demonstrate competence in written English as a prerequisite for their acceptance. One of the measures administered each quarter to the prospective teachers is the Missouri College English Test (Callis & Johnson, 1965), a 90-item multiple-choice test assessing competence in spelling, capitalization, punctuation, sentence style and structure, grammar, and paragraph organization. The

researchers were assigned the task of setting standards for the Missouri test to separate those students at least minimally competent in English from those who are not. The Ebel, Nedelsky, and Angoff methods of standard setting were utilized in this endeavor.

Purpose of the Study

The following research questions were investigated:

1. Do the Ebel, Nedelsky, and Angoff methods of setting standards produce different cutting scores for the same standardized test?
2. Do the same methods, when utilized by raters of differing competency levels, result in different standards for the same test?
3. Are the standards set by the different methods the same across groups of raters, i.e., is there a method-rater group interaction?

Method

Subjects

Subjects were 15 raters that formed three groups of five each. Members of the first group were five advanced doctoral students in English education. All were female with a mean age of 31.6 years and an average of 3.5 years experience teaching English in public schools. They also had taught English at the post-secondary level for an average of 2.4 years. The second group of raters consisted of five high school English teachers employed at a nearby public school when this study was conducted. This group included four females and one male with a mean age of 28.6 years and an average of 4.6 years of high school English teaching experience. All five teachers have their bachelor's degrees in English. Four have obtained their master's degrees while the other has degree work in progress. The third group was composed of five university faculty members with training and/or experience in English education. This group

included four males and one female with a mean age of approximately 44.5 years. The faculty members have an average of 10.8 years teaching experience at the post-secondary level, and all hold terminal degrees.

Procedure

Each rater in each group was asked to rate each of the 90 items on the Missouri College English Test, Form B, according to procedures described by Ebel (1972), Nedelsky (1954), and Angoff (1971). The raters worked independently and were instructed to rate each item all three ways before going to the next item. Detailed rating instructions were provided each rater along with a rating form developed by the researchers and a copy of the Missouri test with the correct responses to each item indicated. Each rater also completed an eight-item personal data questionnaire.

The first item-rating method employed by the raters was Ebel's procedure. On the rating form nine categories were listed, and the rater was instructed to check one of the nine categories indicating whether the Missouri test item being rated was considered essential, important-easy, important-medium, acceptable-easy, acceptable-medium, acceptable-hard, questionable-easy, questionable-medium, or questionable-hard. These categories made it necessary for the rater to decide upon the importance of the item as well as upon the difficulty level of the item according to the judgment of the rater. The reference point for the judgment was the beginning teacher minimally competent in English.

The second method utilized was the Nedelsky approach in which raters were instructed to circle the letter of the option(s) on the Missouri test item that a beginning teacher minimally competent in English should be able to eliminate as incorrect. For this procedure, on the rating form, the letters of the options were printed beside the item number, and the

correct option appeared in a square.

For the third method, described by Angoff, the raters were instructed to enter the percentage of beginning teachers, minimally competent in English, that they thought would respond correctly to the Missouri test item being rated. A slight modification of Angoff's original method was made in that the lowest probability (20% or 25% depending on the number of options) and the highest (100%) were inserted into the rating form. For any proportion different from the minimum or maximum, the raters wrote in the proportional numbers.

After all rating forms were returned, cutting scores for each of the 15 raters and each of the three groups of raters were determined for each of the three methods following procedures described by Ebel, Nedelsky, and Angoff. Cutting scores established by the Ebel procedure were computed using the following table:

Item Category	Number of Items	Expected Success	Number X Success
Essential		1.00	
Important--Easy		.90	
Important--Medium		.70	
Acceptable--Easy		.80	
Acceptable--Medium		.60	
Acceptable--Hard		.40	
Questionable--Easy		.70	
Questionable--Medium		.50	
Questionable--Hard		.30	
	Total = 90* or 450**		Total = Cutting Score

* 90 items for individual raters
**450 items for the combined ratings of all five members in the group.

For each individual the number of items receiving each of the nine respective ratings was multiplied by the expected success rate shown with the total of these resulting values being the cutting score. Group cutting scores were averages of the cutting scores for the individuals within each respective group.

In computing the standard with the Nedelsky procedure, the reciprocals of the number of options remaining after rater elimination of the options minimally competent teachers would recognize as incorrect were summed. These sums served as the cutting scores for individual raters, and the group mean was computed for each of the three groups.

Standards were obtained via the Angoff method by summing the probabilities or proportions established for each item by each rater. The group score represented the mean of the five raters in each group.

To determine if use of the different methods resulted in different standards across groups or if different groups set different standards with the same method, a 3 X 3 analysis of variance was performed with group membership representing the between group factor and rating method the within group factor. Scheffe's multiple range test for mean comparisons was utilized to investigate the presence of significant mean differences.

Results and Conclusions

The analysis of variance shown in Table 1 revealed that no significant differences existed between groups of raters across methods. However, significant differences ($p < .001$) were found between rating methods across groups. There was also a significant interaction between rater groups and methods ($p < .05$).

Insert Table 1 About Here

In Table 2 means are portrayed for all groups and all methods. Row (group across methods) and column (method across groups) means are also

Insert Table 2 About Here

given. A graphic representation of the mean cutting scores set by each group of raters using, respectively, the three rating methods is given in Figure 1.

Insert Figure 1 About Here

As depicted in Table 2, for all rater groups across all three methods the mean cutting scores were relatively stable. However, the differences between methods were not so similar. Scheffe's multiple range test revealed that the difference between Ebel and Nedelsky means and the difference between Angoff and Nedelsky means were significant, $p < .10$ (Scheffe, 1959). The mean cutting score set by the Nedelsky method was significantly lower than the ones set by either the Angoff or the Ebel methods. The Ebel-Angoff comparison did not reach significance.

The group X method interaction was investigated by a Scheffe procedure described by Marascuilo and Levin (1970). This analysis revealed that the source of the interaction stemed from the standard set by the high school teachers using the Nedelsky method as seen clearly in Figure 1. This group established a lower standard via the Nedelsky method.

These findings in summary, as related to the research questions, suggest that:

1. When utilized to set standards on the same test, the Ebel, Nedelsky, and Angoff methods produce different standards.
2. When these methods are combined, groups of raters set standards that do not differ.
3. High school teachers using the Nedelsky procedure set the lowest standards.

Discussion

The findings of Andrew and Hecht (1976), Koffler (1980), and Skakun and Kling (1980) that different rating methods resulted in different cutting scores were substantiated by this study. Such results suggest that extreme caution be exercised when selecting methods for determining cutting scores.

An interesting finding was that when the three judgmental methods investigated in this study were combined, similar standards were set by the different groups. Thus, the utilization of several methods to set proficiency standards is reinforced by the results of this study. Koffler (1980) and Livingston and Zieky (1978) also suggested using more than one standard-setting procedure to determine cut-off scores. Even though a significant method-rater interaction was found, the combination of methods across raters tended to stabilize the cutting scores across groups.

Application of the cutting scores set by each respective method to Missouri test scores of a group of 173 undergraduate students attempting to demonstrate competence in English shows clearly the discrepancies that would result in numbers passed and failed. Use of the Ebel standard set across groups would result in the failure of approximately 58% of the

sample or 100 of the students. Use of the Angoff standard would lead to the failure of 36% or 62 students. Use of the Nedelsky standard would bring about the failure of 12% or 21 students.

Andrew and Hecht (1976) found that use of the Ebel method of standard setting yielded a cutting score equal to 68.4% of the total number of items on their test. In this study use of the Ebel method resulted in a cutting score equal to 69.6% of the items on the Missouri test. In both studies the standard set with the Ebel method was higher than standards set with other methods (i.e., Nedelsky in Andrew & Hecht; Nedelsky & Angoff in the present study). In both studies, too, use of the Ebel method, unlike results with other methods, resulted in similar standards across groups of raters. (See Figure 1 for a graphic depiction of stability in this study.) This stability with the Ebel method across raters may be due to a better understanding on the part of the raters regarding the importance and difficulty of items than their understanding of the process of eliminating options with the Nedelsky procedure or of estimating proportions of people with the Angoff method. Another possible reason for the stability with the Ebel method may be that with it the raters considered two dimensions of the item (importance and difficulty) as opposed to only one respective dimension as considered with the Angoff and Nedelsky methods.

Since with the Ebel method stability across different groups of raters has been demonstrated repeatedly, and since higher standards have been set with the Ebel procedure than with other methods used in conjunction with it, a revised procedure involving lower expected success proportions is suggested that will result in not affecting stability but in lowering the standard set by the Ebel method. In this study the expected success proportions assigned to each category were those given by Ebel

as an example. However, his proportions are flexible and are designed to be determined in advance of the rating process by the judges themselves. By lowering each proportion by 10%, the resulting cut-off score would still be stable across groups but would be lowered. For the Missouri test 53.8 items correct would be the cutting score instead of the 62.6 items originally chosen as the cut-off point. With this new Ebel standard applied to the 173 undergraduates tested for this study sample, only 31% of them would have had scores that fell below the cut-off score. This adjusted procedure for the Ebel method would thus result in a number failing that is no longer substantially discrepant from numbers resulting with use of other methods, and it may be helpful to future users of the Ebel method for setting standards.

Conclusions

Judgmental methods of standard setting, as with other methods, are considered arbitrary procedures (Glass, 1978). Hambleton et al. (1979) recognized this procedural peculiarity. This study has shown that, although arbitrary, when multiple judgmental standard-setting procedures are utilized by different groups concurrently, stability across raters can be achieved and decisions can be made in a relatively judicious manner. A similar result may be effected through the use of a modified Ebel approach.

However, Shepard (1980) stated that no standard setting procedure has been identified that sets a "true" standard. She further stated that competency as an entity to be measured is on a continuum and cannot be dichotomized. Therefore, educators and others faced with the task of setting standards governing competency decisions are encouraged to consider her caution along with the complete findings in this and other related studies before making decisions regarding their standard-setting procedures.

Table 1

Analysis of Variance for Rater Groups and Methods

Source	df	MS	F
Rater Groups	2	13.31	.03
Error (between group)	12	397.40	- - -
Methods	2	1553.67	21.82**
Rater X Method Interaction	4	206.04	2.89*
Error (within group)	24	71.20	- - -

*p < .05.
**p < .001.

Table 2

Means[a] by Groups and Methods

Group	Method			Row Means
	Ebel	Angoff	Nedelsky	
Graduate Students	61.56	51.14	46.61	53.10
High School Teachers	63.94	62.54	34.49	53.32
University Faculty	63.32	54.38	46.80	54.83
Column Means	62.61	56.02	42.63	

[a]Means presented are number of correct items needed to pass the Missouri test (total possible = 90).

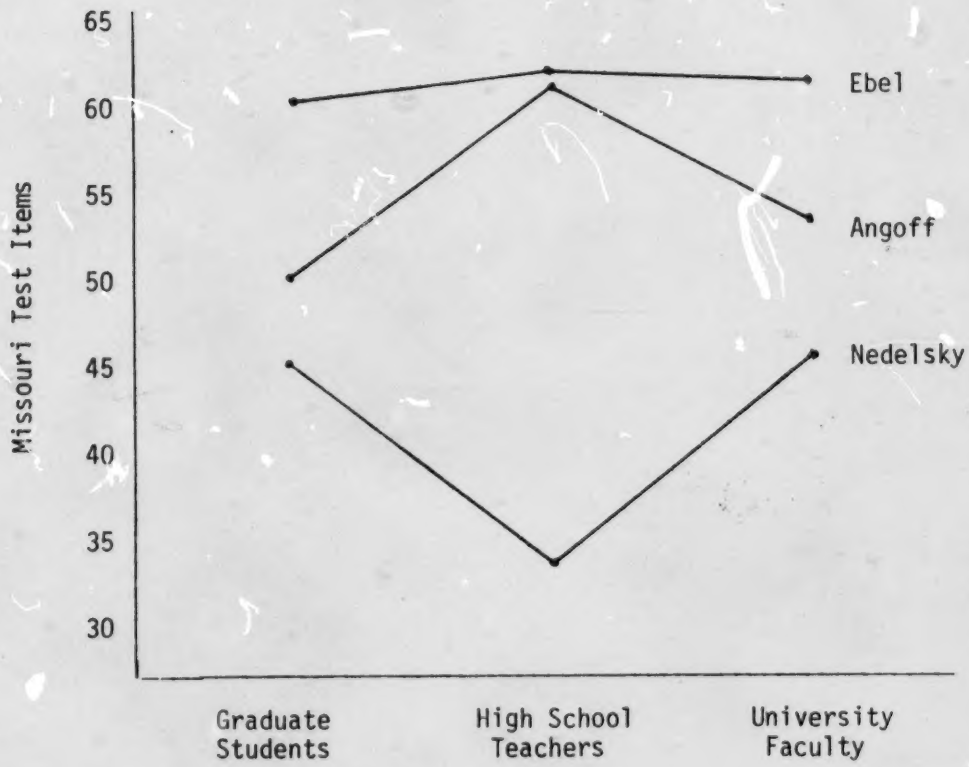

Figure 1. Mean cutting scores for groups by method.

References

Andrew, B. J., & Hecht, J. T. A preliminary investigation of two procedures for setting examination standards. Educational and Psychological Measurement, 1976, 36, 35-50.

Angoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement. Washington, D. C.: American Council on Education, 1971.

Buck, L. S. Guide to the setting of appropriate cutting scores for written tests: A summary of the concerns and procedures. Washington, D. C., Personnel Research and Development Center, United States Civil Service Commission, 1977.

Callis, R., & Johnson, W. Missouri College English Test. New York: Harcourt, Brace, & World, 1965.

Ebel, R. L. Essentials of educational measurement. Englewood Cliffs, N. J.: Prentice-Hall, 1972.

Glass, G. V. Standards and criterion. Journal of Educational Measurement, 1978, 15, 237-261.

Hambleton, R. K. On the use of cut-off scores with criterion-referenced tests in instructional settings. Journal of Educational Measurement, 1978, 15, 277-290.

Hambleton, R. K., Powell, S., & Eignor, D. R. Issues and methods for standard setting. In Laboratory of Pschometric and Evaluative Research Report No. 70. Amherst, MA.: School of Education, University of Massachusetts, 1979.

Koffler, S. L. A comparison of approaches for setting proficiency standards. Journal of Educational Measurement, 1980, 17, 167-178.

Linn, R. L. Demands, cautions, and suggestions for setting standards. Journal of Educational Measurement, 1978, 15, 301-308.

Livingston, S. A., & Zieky, M. J. Manual for setting standards on the basic skills assessment tests. Princeton, N. J.: Educational Testing Service, 1978.

Marascuilo, L. A., & Levin, J. R. Appropriate post hoc comparisons for interaction and nested hypotheses in analysis of variance designs: The elimination of Type IV errors. America Educational Research Journal, 1970, 7, 397-421.

Meskauskas, J. A. Evaluation models for criterion-referenced testing: Views regarding mastery and standard setting. Review of Educational Research, 1976, 46, 133-158.

Nedelsky, L. Absolute grading standards for objective test. Educational and Psychological Measurement, 1954, 14, 3-19.

Popham, W. J. Setting performance standards. Los Angeles, CA.: Instructional Objectives Exchange, 1978.

Scheffé, H. The analysis of variance. New York: John Wiley and Sons, 1959.

Shepard, L. Technical issues in minimum competency testing. In D. C. Berlinger (Ed.), Review of research in education (Vol. 8). Itasca, IL.: F. E. Peacock, 1980.

Skakun, E. N., & Kling, S. Comparability of methods for setting standards. Journal of Educational Measurement, 1980, 17, 229-235.